Graphical Models

Conditional & Local Probability Models

Siamak Ravanbakhsh

Winter 2018
Learning objective

- Conditional Random Fields
- Local probability Models:
  - Deterministic CPDs
  - Noisy-OR model
  - Generalized Linear Model
Conditional Random Fields; **Motivation**

**Structured prediction:** output labels are structured

- $X$ is always observed
- $Y$ is structured

**Examples:**
- *image segmentation*
- *part of speech tagging*
- *optical character recognition*
Conditional Random Fields (CRF)

- a conditional graphical model $P(Y \mid X)$
  - **first attempt:** $P(Y \mid X) = \frac{P(X,Y)}{P(X)}$
- for prediction, no need to model $P(X)$
  - may not have enough data
  - $X$ could be high-dim and complex
Conditional Random Fields (CRF)

second attempt:

\[ P(Y \mid X) = \frac{1}{Z(X)} \tilde{P}(X, Y) = \frac{1}{Z(X)} \prod_k \phi_k(D_k) \]
Conditional Random Fields (CRF)

second attempt:

\[
P(Y \mid X) = \frac{1}{Z(X)} \tilde{P}(X, Y) = \frac{1}{Z(X)} \prod_k \phi_k(D_k)
\]

- differs from MRF in the partition function
  - \textit{input-dependent} \quad Z(X) = \sum_Y \tilde{P}(Y, X)
Conditional Random Fields; a running example

\[ P(Y \mid X) = \frac{1}{Z(X)} \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1}) \]

\[ Z(X) = \sum_Y \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1}) \]

practically the same as

- e.g., in speech recognition (what do potentials encode?)

for each \( X=x \), we have a different MRF
Conditional Random Fields; another benefit

\[
P(Y \mid X) = \frac{1}{Z(X)} \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1})
\]

\[
Z(X) = \sum_Y \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1})
\]

what if \( \phi_i(X, Y_i) \) instead of \( \phi_i(X_i, Y_i) \)?

**sparse structure** after conditioning on \( X=x \)

- learning needs inference on this structure *(discussed later)*
- not true for the corresponding MRF
Conditional Random Fields; input structure

\[ P(Y \mid X) = \frac{1}{Z(X)} \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1}) \]

\[ Z(X) = \sum_Y \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1}) \]

How about the structure of the input?

\[ P(Y \mid X) = \frac{1}{Z(X)} \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1}) \gamma_i(X_i, X_{i+1}) = \]

\[ \frac{1}{Z'(X)} \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1}) \]

i.e., input structure can be ignored
Conditional Random Fields; **parametrisation**

\[
P(Y \mid X) = \frac{1}{Z(X)} \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1})
\]

\[
Z(X) = \sum_Y \prod_{i=1}^{5} \phi_i(X_i, Y_i) \prod_{i=1}^{4} \psi_i(Y_i, Y_{i+1})
\]

- In practice we need to **learn the potentials**
- **Parameterize** them and learn the parameters (e.g., a neural network)
  - traditionally: a log-linear model: \( \phi_i(X_i, Y_i; w_i) \triangleq \exp(\sum_k w_{i,k} f_{i,k}(X_i, Y_i)) \)
    - *E.g.*, for binary input/output:
      \[
      \phi_i(X_i, Y_i; w_i) = \exp(w_i \mathbb{1}(X_i = 1, Y_i = 1)) = \exp(w_i X_i Y_i)
      \]
Local probabilistic models
Local probabilistic models

- conditional probability distributions (CPDs)
  - in prediction $P(Y \mid X_1, \ldots, X_n)$
  - in Bayes-nets $P(X \mid Pa_X)$
    - discrete variables (CPTs)
      - exponential in $|Pa_X|$

- how to represent these efficiently? exploit some sort of structure
Deterministic CPDs

\[ P(X \mid Pa_X) \triangleq \mathbb{I}(X_i = f(Pa_{X_i})) \]

determinism produces **additional independencies**:

- without determinism: \((D \perp E \mid A, B) \notin \mathcal{I}(G)\)
- with determinism: \((D \perp E \mid A, B) \in \mathcal{I}(G)\)
Deterministic CPDs

\[ P(X | Pa_X) \triangleq \mathbb{I}(X_i = f(Pa_{X_i})) \]

determinism produces **additional independencies**:

**without determinism:** \((D \perp E | A, B) \not\in \mathcal{I}(G)\)

**with determinism:** \((D \perp E | A, B) \in \mathcal{I}(G)\)

**deterministic d-separation:** \((X, Y | Z)\)?

- add all the variables that deterministically follow \(Z\) to define \(Z^+\)
- run d-separation for \((X, Y | Z^+)\)
Noisy-OR model

- for **binary** variables only
- number of parameters is linear in $|Pa_{Xi}|$
- each parent ($x_j = 1$) is an **independent cause**
- each cause is observed with prob $P(x_j' = 1) = \lambda_j x_j$

$$p(X_i = 0 \mid Pa_{Xi}) = (1 - \lambda_0) \prod_{X_j \in Pa_{Xi}} (1 - \lambda_j x_j)$$

- leak parameter (role of a bias term) - no cause was observed
Noisy-OR model; **vis.**

\[
p(X_i = 0 \mid Pa_{X_i}) = (1 - \lambda_0) \prod_{X_j \in Pa_{X_i}} (1 - \lambda_j X_j)
\]

- \( \lambda_0 = 0 \)
- \( \lambda_0 = .5 \)

leak parameter (role of a bias term)

no cause was observed
Noisy-OR model; example

Medical diagnosis *(BN2O network)*

\[ p(F_i = 0 \mid Pa_{F_i}) = (1 - \lambda_{i,0}) \prod_{D_j \in Pa_{F_i}} (1 - \lambda_{i,j} D_j) \]

\*CPDs:*
Logistic CPD

for binary output variables

\[ P(X_i = 1) = \frac{\exp(\sum_j w_j X_j)}{1 + \exp(\sum_j w_j X_j)} \]

logistic aggregation function

generally, the input can be **discrete** or **continuous**

- E.g., \( X_j = 2 \) or \( X_j, \ldots, X_{j+n} = 0, 1, \ldots, 0 \)

**one-hot coding**
Logistic CPD

for **binary output** variables

\[
P(X_i = 1) = \frac{\exp(\sum_j w_j X_j)}{1 + \exp(\sum_j w_j X_j)}
\]

logistic aggregation function

generally, the input can be **discrete** or **continuous**

- *E.g.*, \( X_j = 2 \) or \( X_j, \ldots, X_{j+n} = 0, 1, \ldots, 0 \)

one-hot coding

binary input: *each cause has a multiplicative effect on the ratio* \( \frac{P(X_i=1)}{P(X_i=0)} \)
Softmax CPD

extension for categorical outputs
softmax function for aggregation:

\[ f(z_\ell) = \frac{\exp(z_\ell)}{\sum_{\ell'} \exp(z_{\ell'})} \]

functional form of the CPD:

\[ P(X_i = \ell) = \frac{\exp(\sum_j w_{j,\ell} X_j)}{\sum_{\ell'} \exp(\sum_j w_{j,\ell'} X_j)} \]
**Independence of causal influence**

*Commutative and associative aggregation*

**Logistic CPD**
- **transformation**
  
  \[ X'_j = w_j X_j \]

- **aggregation**
  
  logistic function

\[
P(X'_i = 1) \propto w \times \text{active parents} \]

\[
P(X_i = 1) \]

**Noisy-OR**

\[
P(X'_j = 1) = \lambda_j X_j \quad 0 \leq \lambda \leq 1\]

**OR /Max/...**

\[
Pa^g_{X_i}
\]

\[
X_j
\]

\[
X'_j
\]

\[
X_i
\]
**Linear Gaussian CPD**

for **continuous** input/output variables

\[ P(X_i) = \mathcal{N}(\sum_j w_j X_j; \sigma^2) \]

alternatively, a **discrete** input selects among **continuous** coefficients (produces a Gaussian mixture):

\[ P(X_i) = \mathcal{N}(\sum_j w_j, x_d X_j; \sigma_{X_d}^2) \]

**conditional** linear Gaussian CPD:

one Gaussian mixture for each discrete assignment
Generalized linear models

\[ \mathbb{E}[X_i] = f(w^TPa_{X_i}) \]

**mean function**

*Logistic CPD:* *f* is the logistic function

*Gaussian CPD:* *f* is the identity function

\[ X_j \rightarrow \cdots \rightarrow P_{a_{X_i}} \]

\[ X'_j = w_j X_j \]

\[ X_i \]
Generalized linear models

$$E[X_i] = f(w^T P a_{X_i})$$

**Logistic CPD:** $f$ is the logistic function

**Gaussian CPD:** $f$ is the identity function

Conditional dist. is a member of the exponential family

$$p(x_i \mid P a_{X_i}) = h(x_i) \exp(w^T P a_{X_i} - F(w^T P a_{X_i}))$$

(base measure) (integral of $f$)

(will come back to this in exp. family lecture)
Conditional Bayesian networks

use an entire Bayes-net to represent a CPD

\[ P_{\operatorname{a}^g_{X_i}} \text{ is always observed (similar to a CRF)} \]

\[ P(X_i \mid P_{a_{X_i}}) = \sum_Z P(X_i, Z \mid P_{a_{X_i}}) \]
Conditional Bayesian networks; example can be used for encapsulation in complex models.
Summary

the conditioned version of directed & undirected models:

- Conditional Random Fields
- Conditional Bayes-nets

representing conditional probabilities:

- deterministic CPD
- noisy-OR model
- logistic CPD
- linear Gaussian CPD — part of a bigger family of GLMs

how about using neural networks?